Signal Processing: Image Communication 93 (2021) 116179

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

On the performance of objective quality metrics for lightfields

Saeed Mahmoudpour [*,1], Peter Schelkens [1]

Department of Electronics and Informatics, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
imec, Kapeldreef 75, B-3001 Leuven, Belgium

ARTICLE INFO

Keywords:
Lightfield imaging
Quality of experience
Compression
Mean opinion score
Objective quality metrics

ABSTRACT

Lightfield (LF) technology has attained significant attention in recent years due to its capability to capture much richer textural and geometric information in the scene compared to the classical 2D representation. The resampling and compression operations on LFs often lead to visual quality degradation, thus, sophisticated visual quality assessment methods play a crucial role to ensure a pleasant viewing experience. To this end, it is necessary to examine the performance of quality assessment methods for LF contents. The paper provides a comprehensive study on the reliability of various objective algorithms for LF quality prediction. Three subjectively-annotated LF data sets were selected and an extensive quality estimation analysis has been conducted using several objective quality assessment methods. In total, 250 LFs (more than 48000 perspective images) were evaluated. The results were compared against human opinion scores using various correlation indices and their statistical significance. Next, a decision-making strategy was adopted to choose the most reliable quality metrics for evaluation of LFs and finally, a metric fusion framework was proposed to further improve the quality prediction accuracy. To best of our knowledge, the benchmark and the analytical methodologies used in this paper are the most comprehensive study on the objective quality assessment methods for LF application.

1. Introduction

Lightfield (LF) imaging is an emerging technology that enables capturing and simulating visual information at multiple focal distances and positions in the scene. By collecting the radiance of light rays from different directions, LF can benefit from the depth and angular information to provide a richer visual experience that is not possible in traditional photography. The LF representation is defined based on the plenoptic function.

The plenoptic function is an early model of light ray distribution – proposed by Gershun [1] in 1936 and further matured by Adelson and Bergen [2] in 1991 – that gives a comprehensive definition of light rays, with wavelength λ, observed from every position (x, y, z) and direction (θ, ϕ) in free space, and at every time instance t. This demonstration is presented in the form of a 7D full-plenoptic function $\mathcal{L}(\theta, \phi, \lambda, t, x, y, z)$ that is challenging to interpret and utilize for practical applications. Levoy and Hanrahan in 1996 [3] introduced a simplistic lightfield function that demonstrates the plenoptic model over a 4D light ray space in aperture and image planes by quantifying radiance information in spatial (st) and angular (uv) dimensions. A 4D lightfield $\mathcal{L}(u, v, s, t)$ is represented as an array of 2D views arranged in two intersected planes, and it can be captured by assembling an array of cameras (i.e., multi-view) or by using a single microlens camera (*i.e.*, lenslet representation) [4].

The large volume of high dimensional information conveyed in an LF representation urge for a significant increase in the storage demand, transmission bandwidth, and the computational complexity of LF processing. Thus, information handling for efficient LF capturing, rendering and visualizing is a main challenge. To address this problem, LF coding problem has garnered recent attention. Considering the inherent information redundancy of LFs, various processing techniques such as LF data compression and view reconstruction methods have been employed to address the storage demands. The JPEG standardization committee launched a call for proposal on 4D LF compression in 2017 [5] and a standardization activity called JPEG Pleno is carried out to provide a standard framework for coding of the lightfield, point cloud, and holographic contents [6]. A multi-view extensions of H.265/HEVC, MV-HEVC and 3D-HEVC have been defined, extending the temporal prediction modes to the view dimensions [7,8]. In the context of MPEG-I, Coded Representation of Immersive Media, and coupled with the Versatile View Synthesizer (VVS) activity, standardized depth image-based view synthesis software has been specified [9].

LF processing can induce artifacts that disturb visual perception. Therefore, careful consideration should be taken during the design of

* Corresponding author at: imec, Kapeldreef 75, B-3001 Leuven, Belgium.
E-mail address: saeed.mahmoudpour@vub.be (S. Mahmoudpour).
[1] The authors are members of EURASIP.

https://doi.org/10.1016/j.image.2021.116179
Received 26 November 2019; Received in revised form 22 October 2020; Accepted 24 January 2021
Available online 4 February 2021
0923-5965/© 2021 Elsevier B.V. All rights reserved.

LF processing algorithms to maintain the visual quality. To this end, the effects of different visual artifacts introduced during LF processing stages should be further investigated through subjective experiments and objective quality metrics.

Subjective quality assessment has been considered as the most reliable method of measuring the quality of multimedia content while the associated methodologies are costly, tedious and time-consuming. Therefore, they are not suitable for instantaneous feedback in "real-time" applications. Objective quality assessment methods, on the other hand, are the mathematical algorithms capable of predicting visual quality much faster and they can be efficiently deployed in quality-driven optimization frameworks of LF processing chains. The accuracy of an objective quality assessment method is measured based on its consistency with subjective scores as ground-truth data.

Unfortunately, subjective assessment of LF content is not straightforward due to the limitations in LF data acquisition and representation. High-resolution LF capturing requires bulky camera arrays with careful adjustments. As a compact solution, plenoptic LF cameras with arrays of microlenses in front of the sensor have been introduced for consumer applications. In addition to the capturing problems, visual assessment of LF distortions is difficult due to the lack of dedicated LF display devices to enable 6 degree-of-freedom (6DoF) experience. Therefore, current works often simulate the visual experience of LF display using different passive and interactive techniques. Various LF data sets have been released for applications such as depth estimation, compression, and object detection [10–12]. However, due to the aforementioned limitations, only a few data sets with subjectively annotated contents are publicly available for quality assessment studies.

Existing LF data sets for quality assessment purposes differ in their distortion types, scoring methodologies, and data representation methods. Therefore, subjective scores of multiple data sets might not be fully aligned due to the variations in visual properties considered during different subjective tests. For example in the SMART data set [13], LF contents are displayed in the form of 2D still images without considering the depth and temporal distortions, while the MPI-Lightfield data set [14] enables subjects with interactive visualization of all perspective views. Considering different distortion types and methods of LF visualization, it is important to use reliable objective quality metrics that can consistently perform well in different visual conditions and distortion types.

While objective quality assessment methods for LFs are still emerging [15–18], it is important to examine the performance of the current well-established and widely-used metrics for LF quality assessment to provide recommendations to the research community about the most suitable objective metrics to be deployed for evaluation of the new processing algorithms in LF applications. Various methods have been devised so far for visual quality prediction of image and video that are also applied for LF quality assessment. Their accuracy has often been evaluated on only a single data set with a specific subjective scoring strategy [14,19,20]. However, the performance of metrics may vary across multiple data sets with different distortion types and scoring methodologies. In this paper, the goal is to objectively evaluate the quality of LF contents in several data sets and to deploy systematic evaluation methods for choosing the most reliable objective algorithms for quality evaluation of LFs.

The paper provides an extensive benchmark of 24 objective quality metrics including Full-Reference (FR) and No-Reference (NR) image quality assessment (IQA) methods, stereoscopic IQA algorithms, and video quality metrics. We gathered LFs from SMART [13] and MPI-Lightfield [14] data sets as well as a new data set (VALID) [21] established as a part of JPEG-Pleno standardization activity [6,22–24]. The passive subjective tests for the VALID data set were concurrently conducted in our test laboratory at Vrije Universiteit Brussel (VUB) and at École Polytechnique Fédérale de Lausanne (EPFL) test lab.

Using three LF data sets, in total we collected 250 LFs yielding more than 48 000 perspective views for objective quality assessment. The subjective scores of three data sets were aligned using the iterated nested least square algorithm (INLSA) [25] to produce an integrated data set of 250 LFs. All the objective metrics were examined on three selected data sets as well as on the large integrated set. Several evaluation indices were deployed to measure the prediction accuracy, monotonicity, and discriminability (by transforming the quality prediction regression into a classification task) of objective metrics. Moreover, statistical tests were performed for each evaluation index to specify the significance of differences. Considering all the evaluation results and statistical significance tests, a voting strategy was also employed for choosing the most reliable metrics among the analyzed objective methods. Finally, we proposed to combine multiple objective quality algorithms in a single framework using machine learning to achieve higher quality prediction accuracy.

The rest of the paper is organized as follows: Section 2 describes the LF data sets considered in this paper for the quality assessment study. Moreover, the INLSA framework is introduced to align subjective scores and build an integrated data set. Section 3 outlines the extensive analysis of various objective metrics for quality assessment of LF contents. In Section 4, we explain the analysis of Schulze as a decision-making strategy to select the most suitable metrics for LFs. A metric fusion method is demonstrated in Section 5 to further improve the accuracy of the quality prediction. Finally, Section 6 concludes the paper.

2. Subjectively annotated data sets for lightfields

Among the data sets available for LF applications, only a few of them are dedicated to quality assessment purposes. In this work, we selected three public data sets with subjectively annotated LF contents. The data sets encompass various distortion types and subjective methodologies for scoring. The specifications of the three selected data sets are provided below:

2.1. MPI-lightfield data set

Adhikarla *et al.* [14] established a data set of LF images in which both spatial and angular distortions related to light-field representation are considered for quality assessment. The data set includes synthetic and real scenery each represented by 101 perspective views of resolution 960 × 720 with horizontal parallax. For quality evaluation, we focused on the five real-world LFs in this data set (Fig. 1) that are captured using a Canon EOS 5D Mark II camera moving on a linear stage to produce all views.

Three distortion types were considered for real-world images in this data set, occurring during compression, reconstruction, and representation of LFs. The compression artifacts were simulated using the 3D extension of a H.265/HEVC encoder (3D-HEVC) with six selected quantization steps {25, 29, 33, 37, 41, 45}. To simulate reconstruction artifacts, two methods namely nearest neighbor (NN) interpolation and image warping using optical flow estimation (OPT) have been used. These two techniques were employed to generate intermediate views and recover a dense LF from sparse samples in which the distortion levels were defined by adjusting the angular sub-sampling factor (angular resolution). The optical flow (OPT) method has been selected to be included in our experiment. For LF representation artifacts, display crosstalk between views was simulated by applying a Gaussian filter (GAUSS) in angular domain [26]. All three types of distortions (*i.e.* 3D-HEVC [27], OPT, and GAUSS) were provided at six severity levels.

A side-by-side pair-wise comparison (PC) method was used for subjective evaluation in which subjects indicated their preference over two displayed stimuli. The LFs were displayed on an ASUS VG278 27" Full-HD 120 Hz LCD desktop monitor with a pair of 3D active glasses to provide stereoscopic viewing. The motion parallax was reproduced using a head tracking setup enabling subjects to freely explore the LFs for unlimited time and in an interactive manner. A total number of 40 test subjects have participated in the test. To analyze the results of

Fig. 1. Previews of five LF contents in MPI-Lightfield data set.

PC, the raw scores should be first converted to Mean Opinion Score (MOS)-like values. The conversion is commonly performed based on either Bradley–Terry (BT) [28] or Thurstone Case V [29] model. The main difference between the two models is the usage of logistic function to express the quality difference in Thurstone Case V model while the BT model applies a normal distribution [30,31]. In the MPI-Lightfield data set, the subjective values are scaled based on the Thurstone Case V model. The PC subjective scores in this data set are denoted as just-objectionable-difference (JOD) values since the scores imply the difference from a pristine reference LF. Since the PC experiment is inherently relative, JOD values are also relative to their reference image [32], therefore to minimize the content effect, the reference stimuli is always fixed at zero and the JOD of the test stimuli is shifted accordingly [14,33]. The processed JOD values are used in our experiment for comparison with objective quality scores.

From this data set, we examined a total number of 90 LFs (5 contents × 3 distortion types × 6 quality levels) using objective quality assessment metrics.

2.2. SMART data set

Paudyal *et al.* [13] made a data set of sixteen LF images captured using a Lytro Illum camera. The sampled LFs of size 6510 × 9390 were converted to 15 × 15 perspective RGB views of size 625 × 434 for quality evaluation. Some images of the SMART data set include landscape scenes with very small changes between perspective views due to narrow baseline of the capturing device. Here, we selected five contents that can better represent the parallax between perspective views to be included in our experiment. The selected LFs are presented in Fig. 2.

To generate distorted LFs, four compression methods were used in this data set including JPEG, JPEG 2000, H.265/HEVC Intra, and Sparse Set and Disparity Coding (SSDC) [34]. The first three codecs have been selected to be used in our objective evaluation. The LFs were compressed at four quality levels using the three selected codecs. In total, 60 LFs from the SMART data set were chosen for the test.

Nineteen test subjects have participated in the test. The LF contents were rendered in 2D view and refocused at different points to produce all-in-focused view for subjective assessment. Similar to the MPI-Lightfield, a pairwise comparison (PC) method was used for the quality rating in which the user selected the image with higher quality. The test LFs were displayed on a DELL U2413f screen with 1920 × 1200 pixels resolution. The preference scores from PC test are scaled based on the Bradley–Terry (BT) [28,30] model to obtain continuous MOS-like values. To compensate for the relative nature of the BT scores, we assumed the quality score of each reference LF is fixed at zero and the BT scores of test stimuli are computed accordingly.

2.3. VALID data set

A new data set for visual quality assessment of LFs (VALID) was presented in [21]. Five lenslet contents of size 5368 × 7728 pixels were decomposed into 15 × 15 perspective views of size 625 × 434 pixels. The LFs were captured using Lytro Illum camera and then passed through demosaicing and devignetting processes followed by color/gamma corrections to acquire the reference LFs for quality assessment. Fig. 3 shows the center view of the lenslet images in the VALID data set.

The selected LFs were compressed using five different codecs including H.265/HEVC, VP9, and three state-of-the-art algorithms (P3 [35], P4 [36], and P5 [37]) proposed for LF compression. For H.265/HEVC and VP9 encoding, the perspective views were arranged in a temporal serpentine order to produce pseudo-video sequence. The distortions (quality levels) were provided by encoding at four different bitrates: 0.75 bpp (bit-per-pixel), 0.1 bpp, 0.02 bpp, and 0.005 bpp.

A passive method was employed to compare the LF compression techniques. In this passive approach, perspective views were presented in serpentine order in the form of a pseudo-video at 10 frames per second to the users. The pair-wise test was set up according to ITU-R Recommendation BT.500-13 based on a 7-point grading, from −3 (much worse) to +3 (much better), 0 denoting no preference between a compressed and reference LF. The LFs were displayed on Eizo ColorEdge CG318-4K screen to a total number of 28 subjects. These subjective tests were collaboratively conducted at VUB and EPFL test laboratories and the collected MOS values were successfully cross-checked for consistency. Since the results obtained at VUB and EPFL labs were highly correlated [38], here we solely use the results of subjective testing at VUB for objective quality comparison.

Considering five LFs compressed at four quality levels using five different codecs, a total number of 100 LF images have been evaluated in this data set.

This section outlined the three data sets used for our objective quality assessment. Table 1 summarizes the details of the data sets, subjective test properties, and the number of selected LFs used in our experiment.

2.4. Integrated data set with aligned subjective scores

The selected data sets follow different methods to visualize LFs for quality ratings. In the VALID test, the subjects observed pseudo-video sequences within a limited period while in MPI-Lightfield test, the subjects are allowed to freely explore the content in different viewpoints for an unlimited time. The SMART test relies on the visualization of a single all-in-focus image while the other two data sets are following multi-view observation methods. Moreover, data sets utilize different test material, display devices, and scoring scales. Therefore, MOS values of the different data sets may not be suitably aligned by linear mapping. Fig. 4(a) indicates the original subjective scores (with no correction) versus the Gradient Magnitude Similarity Deviation (GMSD) objective metric [39]. In Fig. 4(b), the subjective scores of each data set is linearly-rescaled in order to map the three score sets to an identical range. As shown in Fig. 4(b), a similar level of distortion (in terms of GMSD) corresponds to different subjective scores in three data sets.

In Fig. 4(c), we deployed the iterative nested least square algorithm (INLSA) [25] to align subjective scores of the three data sets. INLSA maps the scores of multiple data sets to a single scale using multiple well-correlated objective scores. Initially, the algorithm normalizes the subjective scores of the input data sets to the [0, 1] interval. Next, the scores are aligned by solving two least-square (LS) problems. The first LS problem operates on each single data set separately and aims

Fig. 2. Previews of five LF contents selected from SMART data set.

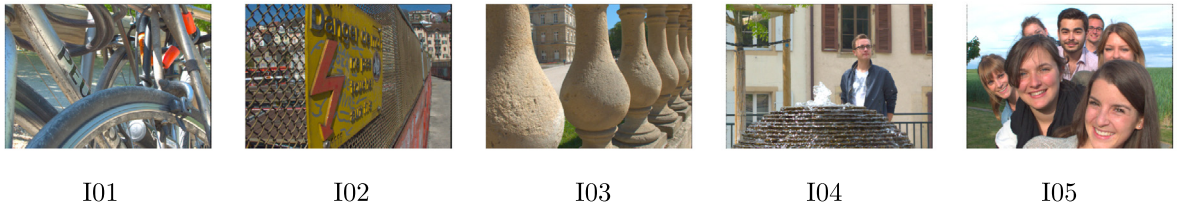


Fig. 3. Central perspective view of five contents used in VALID data set.

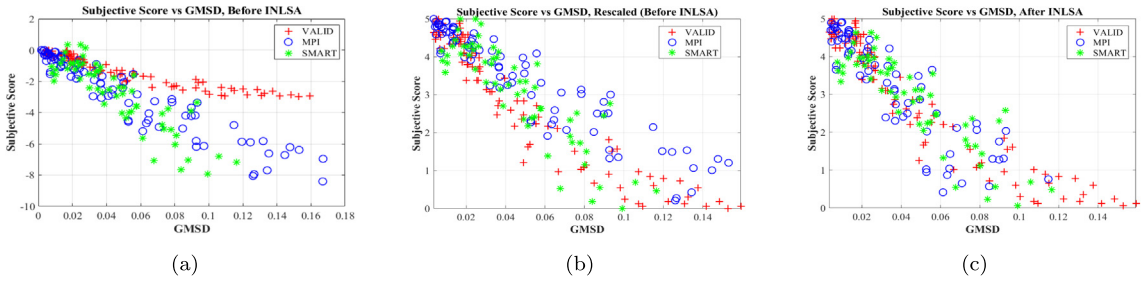


Fig. 4. Integration of the three data sets using INLSA algorithm. (a) GMSD vs. subjective (no correction); (b) GMSD vs. subjective (linear rescaling per data set); (c) GMSD vs. subjective (INLSA).

Table 1
Selected specifications from three data sets used in our evaluation.

Data set Name	Resolution	Subj. method	Dist. Types	Display Type	Total No. selected LFs
MPI-Lightfield	(960 × 720 × 101)	DSIS (Interactive)	3D-HEVC, OPT, GAUSS	ASUS VG278 2700	90
SMART	(625 × 434 × 15 × 15)	DSIS (All-in-focused view)	JPEG, JPEG2000, H.265/HEVC Intra	DELL U2413f	60
VALID	(625 × 434 × 15 × 15)	DSIS (Passive)	H.265/HEVC, VP9, P3, P4, P5	Eizo ColorEdge CG318-4K	100

to bring all scores of each subjective test into the same scale by providing a single first-order correction. The second LS problem finds an optimal combination of objective weights across all data sets such that a subjective score can be predicted by linear combination of the objective scores. Finally, the INLSA delivers optimum model weights and scaling factors by iterating over these two LS problems. The objective methods that are used for INLSA should be highly correlated with subjective values. Here, we selected ten objective metrics for score alignment that have higher performance on three test data sets, including the Feature SIMilarity index (FSIM) [40], GMSD [39], Visual Information Fidelity (VIF) [41], Multi-Scale SSIM (MS-SSIM) [42], Most Apparent Distortion (MAD) [43], Sparseness Significance Ranking Measure (SSRM) [44], Vis3 [45], High Dynamic Range Visual Difference Predictor (HDR-VDP-3) [46], Information content Weighted PSNR (IW-PSNR) and Information content Weighted Structural SIMilarity measure (IW-SSIM) [47].

Fig. 4c shows the scatter plot of the aligned subjective scores versus the GMSD values after applying INLSA. As shown in the figure, the distribution of subjective scores in three data sets are better aligned and data points with close GMSD scores are more likely to have rather a similar subjective quality. Using INLSA, the subjective scores from three data sets are aligned in a coherent way to make a large integrated data set containing 250 LFs for objective metric evaluation task.

3. Performance analysis of the objective quality metrics

The quality of the selected LFs was evaluated using different objective quality assessment algorithms. Then, the objective scores were compared against subjective values in the three data sets as well as the integrated one. To this end, several evaluation criteria were deployed to provide an in-depth assessment of the objective quality metrics.

3.1. Objective quality assessment methods

Different types of objective methods have been utilized for quality assessment in our experiment. In addition to the commonly used Full-reference (FR) image quality metrics (IQMs), we employed several No-reference (NR) IQMs, stereoscopic IQMs, and video quality metrics. Thirteen FR algorithms have been used including the Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity index (SSIM) [48], Multi-Scale SSIM (MS-SSIM) [42], Information Fidelity Criterion (IFC) [49],

Table 2
LCC between the results of each objective metric and subjective scores in three LF data sets.

	VALID		MPI-Lightfield		SMART		Integrated	
	Linear	Logistic	Linear	Logistic	Linear	Logistic	Linear	Logistic
PSNR	0.902	0.951	0.881	0.914	0.710	0.775	0.832	0.875
SSIM	0.930	0.964	0.925	0.932	0.800	0.816	0.783	0.816
MS-SSIM	0.878	0.972	0.890	0.943	0.848	0.860	0.861	0.895
IFC	0.928	**0.983**	0.736	0.795	0.654	0.744	0.789	0.851
UQI	0.937	0.955	0.807	0.846	0.717	0.772	0.811	0.819
VSNR	0.891	0.919	0.850	0.928	0.507	0.559	0.790	0.833
VIF	**0.971**	0.979	0.886	0.896	0.700	0.764	0.879	0.892
IW-PSNR	0.932	0.967	0.885	0.916	0.750	0.837	0.882	0.915
IW-SSIM	0.890	0.978	0.870	0.944	0.846	0.857	0.873	0.925
GMSD	0.944	0.972	0.943	0.949	0.873	0.886	**0.931**	**0.942**
HDR-VDP 3.0	0.948	0.959	0.941	0.944	0.796	0.853	0.881	0.881
MAD	0.944	0.961	**0.956**	**0.960**	0.860	**0.923**	0.921	0.937
SSRM	0.954	0.974	0.928	0.948	0.727	0.744	0.862	0.893
FSIM	0.911	0.977	0.887	0.943	0.875	0.876	0.846	0.899
NIQE	0.548	0.556	0.434	0.604	0.167	0.282	0.132	0.223
BRISQUE	0.827	0.877	0.505	0.515	0.05	0.223	0.539	0.573
IL-NIQE	0.858	0.888	0.529	0.537	0.165	0.301	0.451	0.464
dipIQ	0.791	0.833	0.508	0.612	0.401	0.481	0.574	0.612
MEON	0.804	0.826	0.549	0.651	0.280	0.289	0.527	0.595
HOSA	0.864	0.868	0.600	0.627	0.461	0.568	0.678	0.686
MJ3DFR	0.882	0.970	0.876	0.930	0.843	0.853	0.839	0.874
FI-PSNR	0.683	0.706	0.778	0.851	0.728	0.828	0.703	0.741
STMAD	0.778	0.802	0.857	0.894	0.625	0.786	0.760	0.793
Vis3	0.897	0.950	0.912	0.946	**0.879**	0.892	0.891	0.911

Universal Quality Index (UQI) [50], Visual Signal-to-Noise Ratio (VSNR) [51], VIF [41], IW-PSNR, IW-SSIM [47], GMSD [39], MAD [43], SSRM [44], and FSIM [40]. NR methods are essential in many applications where the reference LFs are not available for quality comparison. The selected NR methods are Naturalness Image Quality Evaluator (NIQE) [52], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [53], Integrated Local NIQE (IL-NIQE) [54], quality-Discriminable Image Pairs inferred Quality (dipIQ) [55], Multi-task End-to-end Optimized deep neural Network (MEON) [56], and High Order Statistics Aggregation (HOSA) [57]. The objective metrics were applied to evaluate the quality of every perspective view and the final scores were obtained by average pooling. Considering 255 views per LF in VALID and SMART data sets and 101 views per MPI-lightfield content, a total number of 48 090 images were evaluated using each objective IQA method. To consider binocular properties, we used two stereoscopic IQA metrics namely MJ3DFR [58] and Frequency-Integrated (FI)-PSNR [59] for quality assessment in which the averaged score is obtained from all stereopairs (*i.e.* adjacent perspective views). We also employed two video quality metrics including Vis3 [45], and Spatio-Temporal Most-Apparent-Distortion model (STMAD) [60]. For video quality assessment, pseudo-video files are generated at 10 fps by arranging the views in linear order for MPI-Lightfield and spiral order for VALID and SMART data sets.

3.2. Performance evaluation

The scores from each objective method are compared against subjective values (as ground-truth data) in terms of accuracy, monotonicity, and consistency of predictions using four evaluation criteria: Linear Correlation Coefficient (LCC), Spearman Rank-order Correlation Coefficient (SRCC), Root Mean Square Error (RMSE), and Outlier Ratio (OR) [61,62]. LCC and RMSE denote the prediction accuracy, SRCC measures the prediction monotonicity, and the OR index quantifies the ratio of predicted scores that are not in the confidence interval (CI) of subjective scores (considered as outliers) and it is utilized to quantify the consistency of predictions. The evaluation indices are summarized as follows:

LCC and SRCC are bounded in [0, 1] in which a higher value means stronger correlation. Lower values of RMSE and OR imply higher performance. We fit a commonly used logistic regression as mapping function between the objective and subjective scores:

$$F(q) = \gamma_1 + \frac{\gamma_2 - \gamma_1}{1 + e^{\frac{-(q-\gamma_3)}{\gamma_4}}} \tag{1}$$

where γ_1 to γ_4 denote the fitting parameters. The mapping function finds a non-linear relation between objective and subjective scores by minimizing the least-square errors. Tables 2 to 5 present the performance comparison results in terms of LCC, SRCC, RMSE, and OR, respectively for all four data sets.

The results show that the performance of the objective metrics is different across data sets because of the variety of content characteristics and distortion types. It is observed that there is not a single metric that can outperform all other methods across all four data sets. The best metrics per data set is always selected from one of the FR methods. Among the NR methods that have lower performance than other algorithms, IL-NIQE and HOSA outperform in the VALID data set, while all NR metrics show weak performance on other three data sets. The stereoscopic 3D metric, MJ3DFR, and the video quality metric, Vis3, consider binocular and temporal distortions, respectively, and they indicate reasonably good performance on all data sets. However, they cannot outperform some FR IQA methods.

Despite of the observed performance variations of the quality metrics, there are some good metrics with rather consistent performance across different data sets. For example, MAD and GMSD do not have the best performance over all data sets; however, they are always one of the well-performing metrics and have been selected as the best metrics in the integrated data set due to their consistently high performance.

3.3. Classification accuracy

Since the subjective scores are random variables with assigned uncertainty, some of these values might not be significantly different from each other, *i.e.* accounting for rather same image quality. In other words, two images with the same visual quality may be assigned with various subjective values while the difference is insignificant (statistical equivalence). Here, we employed an evaluation method proposed by Krasula et al. [63] that considers whether two subjective scores are sufficiently different to account for different visual quality. The basic

Table 3
SRCC between the results of each objective metric and subjective scores in three LF data sets.

	VALID	MPI-Lightfield	SMART	Integrated data set
PSNR	0.932	0.929	0.752	0.883
SSIM	0.950	0.912	0.785	0.834
MS-SSIM	0.962	0.919	0.810	0.895
IFC	**0.966**	0.725	0.696	0.820
UQI	0.931	0.740	0.695	0.785
VSNR	0.914	0.938	0.507	0.836
VIF	0.962	0.860	0.695	0.875
IW-PSNR	0.947	0.930	0.785	0.906
IW-SSIM	0.965	0.932	0.793	0.915
GMSD	0.954	0.945	**0.858**	**0.934**
HDR-VDP 3.0	0.951	0.916	0.798	0.878
MAD	0.934	**0.950**	0.854	0.908
SSRM	0.958	0.944	0.692	0.887
FSIM	0.963	0.933	0.831	0.889
NIQE	0.458	0.165	0.134	0.194
BRISQUE	0.840	0.416	0.024	0.468
IL-NIQE	0.839	0.410	0.151	0.427
dipIQ	0.768	0.268	0.460	0.551
MEON	0.814	0.439	0.320	0.470
HOSA	0.840	0.504	0.439	0.641
MJ3DFR	0.956	0.872	0.828	0.863
FI-PSNR	0.706	0.874	0.775	0.750
STMAD	0.794	0.865	0.677	0.777
Vis3	0.908	0.948	0.841	0.896

Table 4
RMSE between the results of each objective metric and subjective scores in three LF data sets.

	VALID	MPI-Lightfield	SMART	Integrated data set
PSNR	0.297	0.940	1.352	0.797
SSIM	0.255	0.837	1.236	0.952
MS-SSIM	0.227	0.766	1.090	0.721
IFC	**0.175**	1.404	1.428	0.865
UQI	0.286	1.233	1.359	0.944
VSNR	0.379	0.859	1.774	0.911
VIF	0.194	1.028	1.380	0.745
IW-PSNR	0.246	0.929	1.170	0.662
IW-SSIM	0.201	0.764	1.102	0.623
GMSD	0.225	0.731	0.990	**0.551**
HDR-VDP 3.0	0.273	0.764	1.110	0.748
MAD	0.265	**0.645**	**0.819**	0.573
SSRM	0.217	0.737	1.430	0.741
FSIM	0.205	0.766	1.030	0.721
NIQE	0.801	1.844	2.053	1.605
BRISQUE	0.464	1.982	2.083	1.349
IL-NIQE	0.444	1.951	2.041	1.458
dipIQ	0.533	1.829	2.034	1.302
MEON	0.544	1.756	2.049	1.323
HOSA	0.479	1.801	1.761	1.198
MJ3DFR	0.236	0.847	1.117	0.748
FI-PSNR	0.683	1.213	1.198	1.105
STMAD	0.575	1.038	1.321	1.003
Vis3	0.301	0.751	0.964	0.678

idea of the method is to transform the conventional method of accurate prediction into a binary classification problem.

For each pair of stimuli in the data set, the first step is to determine whether a pair of the corresponding subjective scores are significantly different or not. Here, we applied the z-test pairwise statistical comparison as follows:

$$z(i,j) = \frac{|\mu_i - \mu_j|}{\sqrt{\frac{s_i}{N_i} + \frac{s_j}{N_j}}} \tag{2}$$

where μ and s are the mean subjective score and variance of the scores, and N is the total number of subjects. Then, a z-value inside the critical region rejects the null hypothesis H_o (no significant difference) in favor of H_1 (there is a significant difference) at significance level of 5%.

Table 5
OR between the results of each objective metric and subjective scores in three LF data sets.

	VALID	MPI-Lightfield	SMART	Integrated data set
PSNR	0.430	0.822	0.766	0.720
SSIM	0.390	0.867	0.817	0.748
MS-SSIM	0.350	0.800	0.783	0.680
IFC	0.240	0.900	0.866	0.716
UQI	0.430	0.866	0.750	0.828
VSNR	0.460	0.822	0.850	0.804
VIF	**0.230**	0.833	0.833	0.692
IW-PSNR	0.360	0.855	0.766	0.692
IW-SSIM	0.260	0.777	0.733	0.684
GMSD	0.390	0.800	0.666	**0.632**
HDR-VDP 3.0	0.440	0.789	0.716	0.712
MAD	0.400	**0.733**	**0.533**	0.700
SSRM	0.310	0.811	0.866	0.768
FSIM	0.290	0.800	0.766	0.704
NIQE	0.820	0.922	0.916	0.928
BRISQUE	0.650	0.933	0.850	0.892
IL-NIQE	0.560	0.911	0.816	0.896
dipIQ	0.730	0.922	0.916	0.880
MEON	0.670	0.878	0.900	0.880
HOSA	0.580	0.866	0.950	0.832
MJ3DFR	0.410	0.855	0.733	0.772
FI-PSNR	0.700	0.844	0.816	0.804
STMAD	0.600	0.788	0.783	0.800
Vis3	0.520	0.811	0.716	0.748

The same procedure was performed for subjective score pairs in all the data sets to determine binary classes of "significantly different" and "similar pairs". Fig. 5 shows the binary classification result of subjective pairs in the integrated data set. The values are sorted from lowest to highest in which white and black values imply "similar" and "significantly different" pairs (classes), respectively.

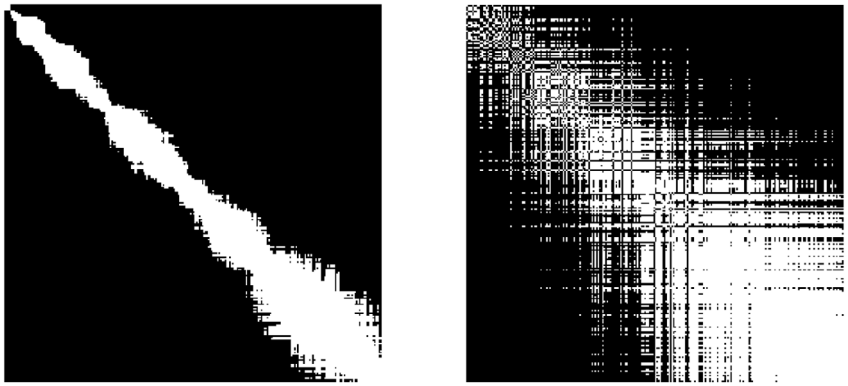
To classify the objective scores, the difference between each pair of the predicted scores is computed as $\delta_{ij} = |q_i - q_j|$ and compared against a threshold Th. The classification performance depends on the selected Th value, thus by changing the threshold value, Receiver Operating Characteristic (ROC) curves can be generated and the Area Under Curve (AUC) denotes the classifier accuracy. The AUC denotes the ability of a metric for discrimination between images of the same and different quality. Table 6 presents the AUC values for each objective metric in which GMSD, MAD, and IW-SSIM are the best metrics in the integrated data set in terms of discriminability. In addition to the AUC index, the balanced classification accuracy is computed as in [64,65]:

$$Acc = 2\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{3}$$

where True Positive (TP) is the number of images with different subjective scores correctly classified as being different by the objective metric and False Positive (FP) is the number of images wrongly classified as being of different quality while they have similar subjective scores. The maximum accuracy achieved by varying the Th value is reported in Table 6. The estimated equivalence matrix of GMSD, with Th value adjusted for the highest accuracy, is depicted in Fig. 5.

3.4. Statistical significance test

To understand the significance of differences between the performance of the objective methods, statistical significance tests were performed according to ITU-T Recommendation P.1401 using the evaluation indices. The significance test methods for each index are briefly discussed in the following paragraphs.



(a) (b)

Fig. 5. Binary classification matrix. (a) Subjective score classification matrix at 95% confidence level. (b) classification matrix for GMSD algorithm (The threshold parameter is adjusted for maximum accuracy).

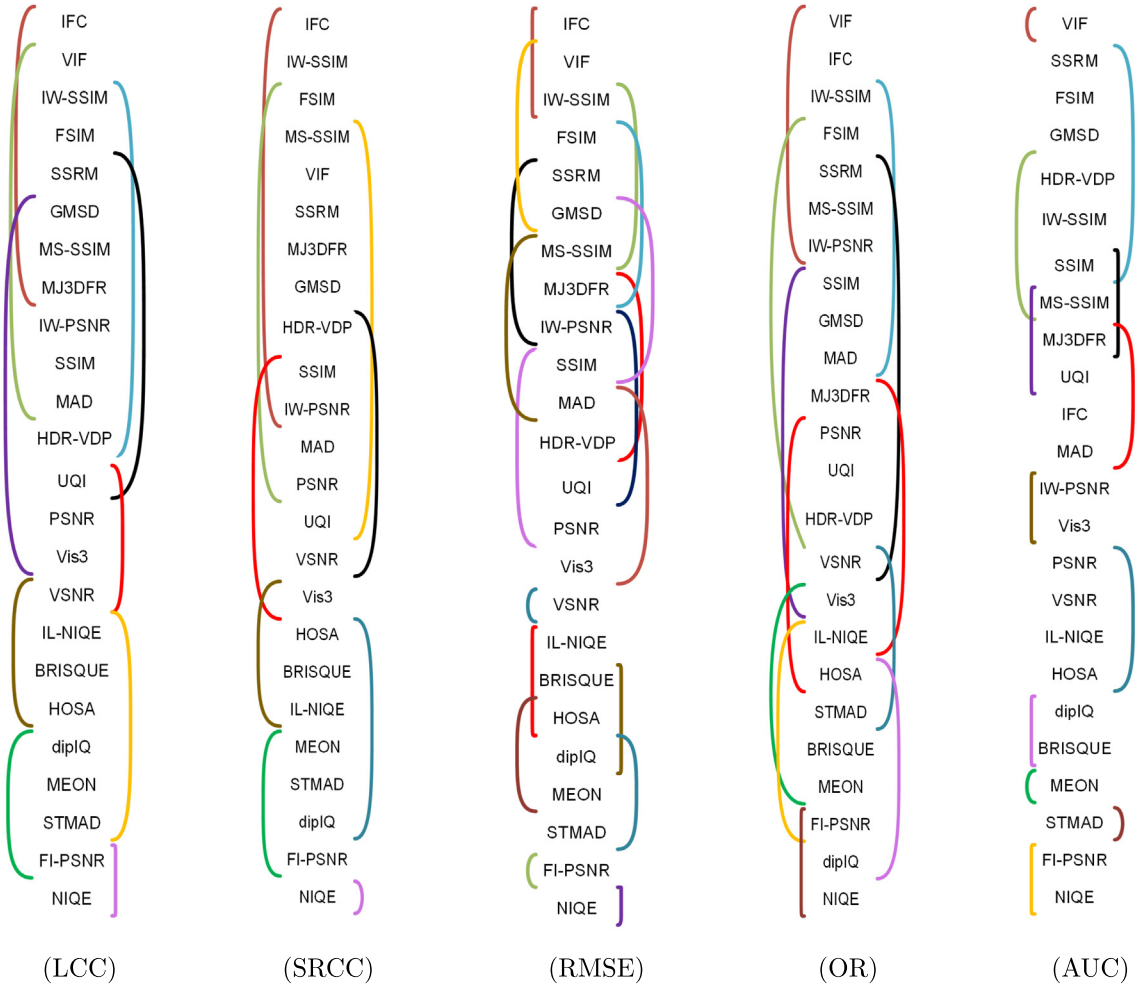


Fig. 6. Statistical significance test on the VALID data set.

3.4.1. LCC and SRCC

Assuming that the subjective and predicted scores (obtained by (4)) follow a bivariate normal distribution, the sampling distribution of LCC is transformed using the Fisher z-transformation so that the transformed distribution $F(LCC)$ becomes normally distributed:

$$z = F(LCC) = 0.5\ ln\frac{1+LCC}{1-LCC} \tag{4}$$

and the standard deviation is computed as:

$$\sigma_z = \sqrt{\frac{1}{N-3}} \tag{5}$$

where N is the number of samples. A two-sample test is employed to evaluate the statistical significance of differences between LCC values of two objective metrics. The null hypothesis is that there is no significant difference and it is rejected when the difference is statistically

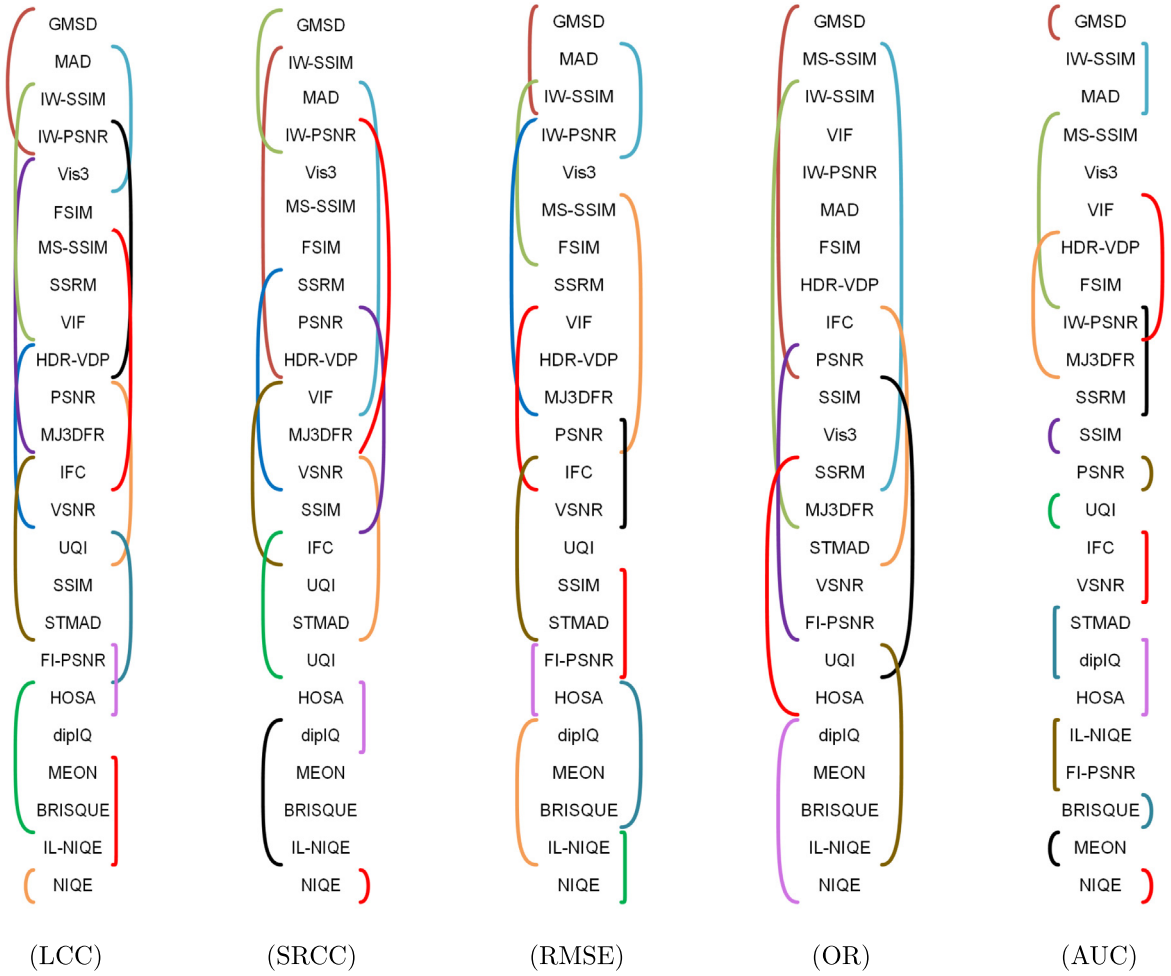


Fig. 7. Statistical significance test on the integrated data set.

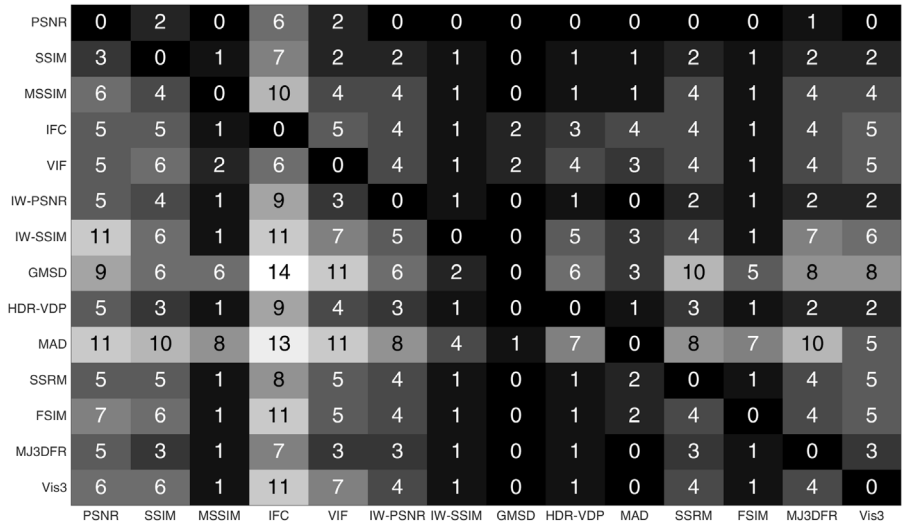


Fig. 8. Pairwise comparison of the strength values for 14 best quality assessment methods.

significant. The observed z-value (z_o) is computed as follows:

$$z_o = \frac{z_1 - z_2 - \mu_{z_1-z_2}}{\sigma_{z_1-z_2}} \tag{6}$$

where $\sigma_{z_1-z_2} = \sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2}$ and $\mu_{z_1-z_2}$ is zero due to the null hypothesis. The observed z-value is checked if it is inside the critical region (determined by 95% two-tailed z-value) to reject the null hypothesis. Same method of significance testing was applied for SRCC values.

3.4.2. RMSE

Assuming a normal distribution of subjective and predicted scores, the distribution of RMSE values follows an approximately chi-squared

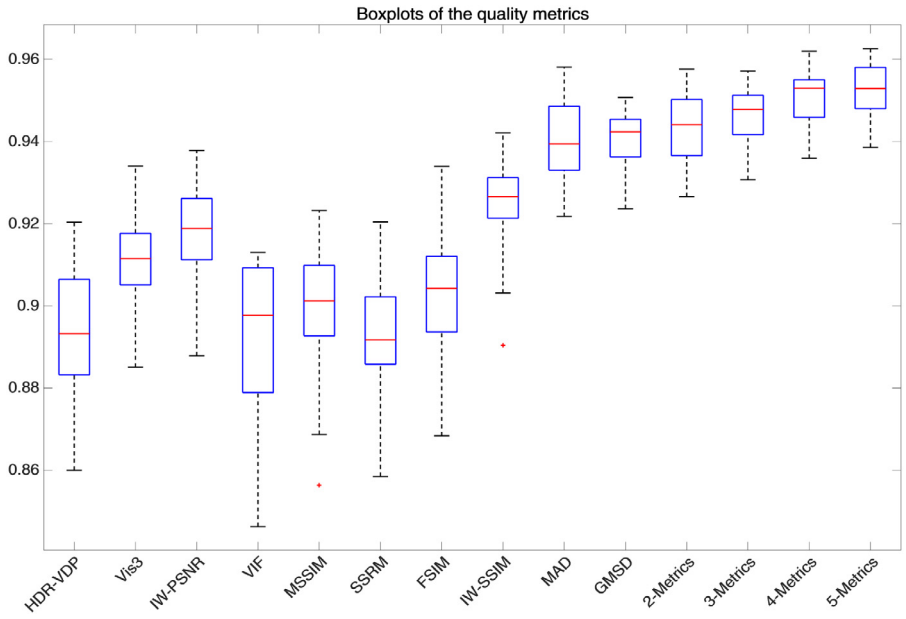


Fig. 9. Boxplots of the single and combined metrics in terms of LCC.

Table 6
AUC and the accuracy of the objective metric on the LF data sets.

	VALID		MPI-Lightfield		SMART		Integrated	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
PSNR	0.785	0.727	0.794	0.751	0.600	0.609	0.726	0.674
SSIM	0.889	0.826	0.792	0.735	0.708	0.682	0.743	0.696
MS-SSIM	0.878	0.823	0.784	0.730	0.732	0.694	0.783	0.720
IFC	0.860	0.780	0.640	0.626	0.596	0.605	0.688	0.639
UQI	0.862	0.800	0.701	0.670	0.658	0.633	0.708	0.671
VSNR	0.782	0.732	0.773	0.736	0.522	0.573	0.684	0.650
VIF	**0.921**	**0.857**	0.757	0.719	0.630	0.623	0.778	0.723
IW-PSNR	0.836	0.774	0.797	0.759	0.635	0.643	0.767	0.715
IW-SSIM	0.891	0.837	0.776	0.734	0.735	0.699	0.805	0.735
GMSD	0.895	0.832	0.822	0.760	0.748	0.710	**0.830**	**0.771**
HDR-VDP 3.0	0.892	0.831	0.816	0.761	0.691	0.667	0.776	0.719
MAD	0.857	0.799	**0.849**	**0.805**	0.778	0.728	0.802	0.744
SSRM	0.904	0.837	0.808	0.755	0.629	0.619	0.760	0.712
FSIM	0.896	0.834	0.772	0.724	0.738	0.695	0.776	0.721
NIQE	0.566	0.564	0.583	0.600	0.514	0.566	0.541	0.546
BRISQUE	0.708	0.663	0.574	0.588	0.529	0.574	0.585	0.570
IL-NIQE	0.768	0.725	0.572	0.584	0.557	0.590	0.603	0.588
dipIQ	0.726	0.688	0.610	0.613	0.550	0.598	0.626	0.603
MEON	0.680	0.648	0.588	0.592	0.543	0.572	0.559	0.551
HOSA	0.768	0.708	0.614	0.617	0.556	0.590	0.617	0.599
MJ3DFR	0.872	0.818	0.756	0.717	0.729	0.688	0.765	0.713
FI-PSNR	0.577	0.565	0.675	0.654	0.627	0.640	0.599	0.582
STMAD	0.647	0.615	0.731	0.695	0.559	0.597	0.637	0.607
Vis3	0.822	0.789	0.808	0.763	**0.784**	**0.737**	0.783	0.735

distribution with a degree of freedom equal to $N-d$, where d denotes the degrees of freedom for fitting function. The null hypothesis – *i.e.* the difference between RMSE of two objective metrics is not significant – is rejected when there is a significant difference. The following statistic has an F-distribution:

$$F_o = \frac{RMSE_1^{\,2}}{RMSE_2^{\,2}} \tag{7}$$

An observed F-value inside the critical region (determined by 95% two-tailed F-value) infers the rejection of the null hypothesis at a 5% significance level.

3.4.3. OR

The OR values are characterized by a binomial distribution with mean $p = OR$ and standard deviation $\sigma_p = \sqrt{\frac{p(1-p)}{N}}$. The statistical difference between OR values are determined by a two-sample test based on the null hypothesis, *i.e.* there is no significant difference between computed OR values. The distribution of differences of two binomially distributed samples follow approximately a normal distribution with normalized z-value calculated as:

$$z_n = \frac{p_1 - p_2 - \mu_{p_1-p_2}}{\sigma_{p_1-p_2}} \tag{8}$$

Due to the null hypothesis assumption, the mean is zero and the standard deviation is computed as:

$$\sigma_{p_1-p_2} = \sqrt{\frac{2p(1-p)}{M}} \tag{9}$$

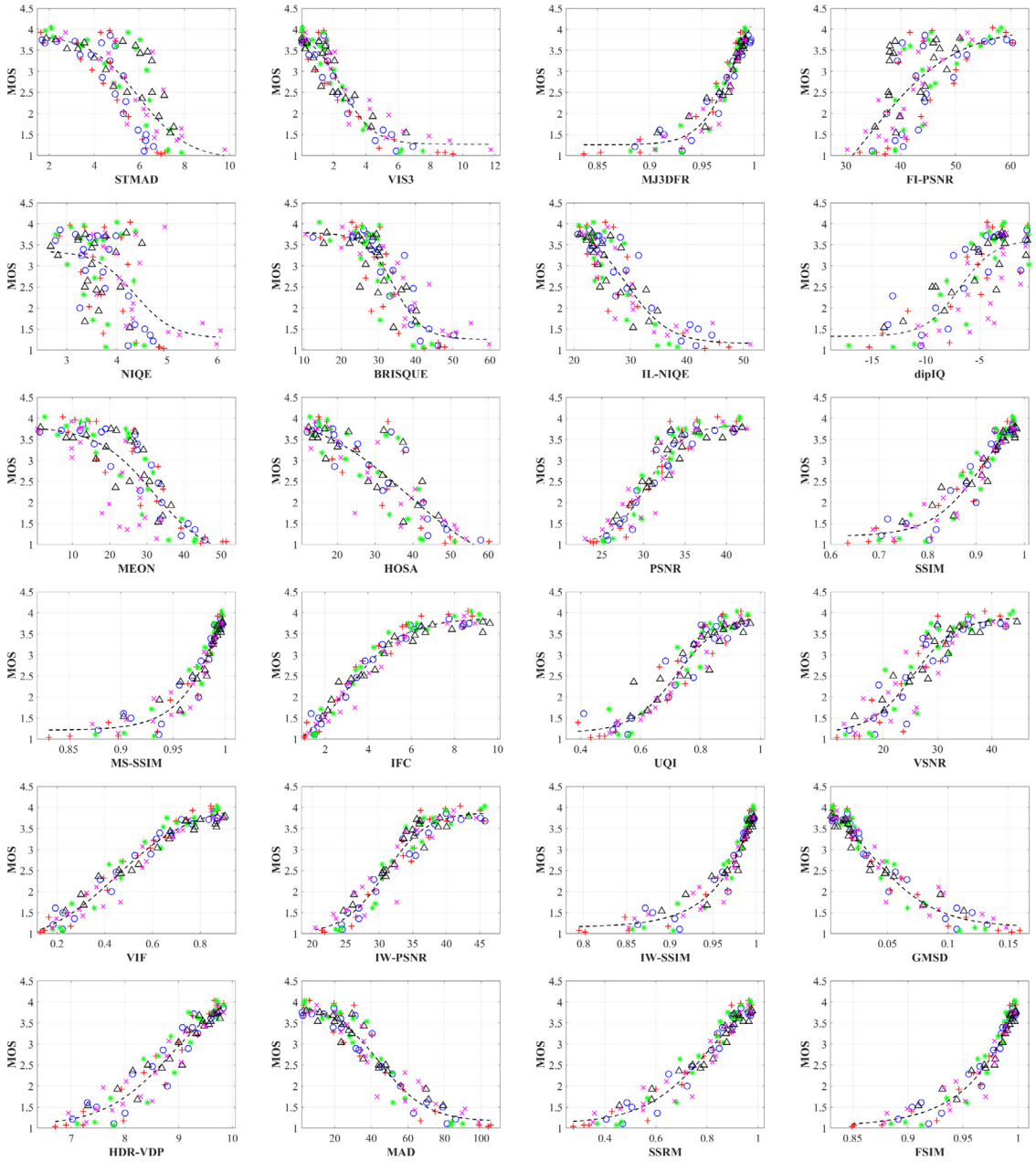


Fig. 10. Scatter plots of Objective metrics vs. MOS (VALID).

The observed z_n value inside the critical region rejects the null hypothesis at a 5% significance level.

3.4.4. AUC

A statistical difference test between AUC values from ROC analysis is performed based on the method proposed in [63]. A critical value c_{ij} between AUC values is obtained as follows:

$$c_{ij} = \frac{AUC_i - AUC_j}{\sqrt{e_i^2 + e_j^2 - 2\rho e_i e_j}} \tag{10}$$

where e is the standard error and ρ is the estimated correlation between two AUC values provided in the table in [66]. Then, the probability of the difference is computed from the cumulative distribution function (CDF) of the normal distribution in which the probability higher than the significance level of 95% accounts for statistical difference.

The statistical significance test was performed on all four data sets. Figs. 6 and 7 present the results of the test on the VALID and the integrated data sets. The objective methods are sorted in descending order in terms of five evaluation indices. The algorithms with the same level of significance (statistical equivalence) are separated in parentheses. In the VALID data set, several objective methods are on top of the list which have statistically equivalent performance in terms of LCC, SRCC, and OR. However, only a few metrics are significantly better in terms of RMSE and AUC. According to the AUC results in the integrated data set, GMSD is statistically better than all other approaches. However, the significance of superiority of this metric is challenged by the other four evaluation indices. According to LCC, SRCC, and RMSE indices, the GMSD has statistically equivalent performance compared to the several methods such as IW-SSIM and MAD.

The evaluation results provide a general overview of the effective metrics for the LF application. However, the decision about the most reliable method is rather challenging considering all performance variations across the four data sets. In the next section, we proposed to use a decision-making strategy to provide better clarification about the best objective quality assessment methods.

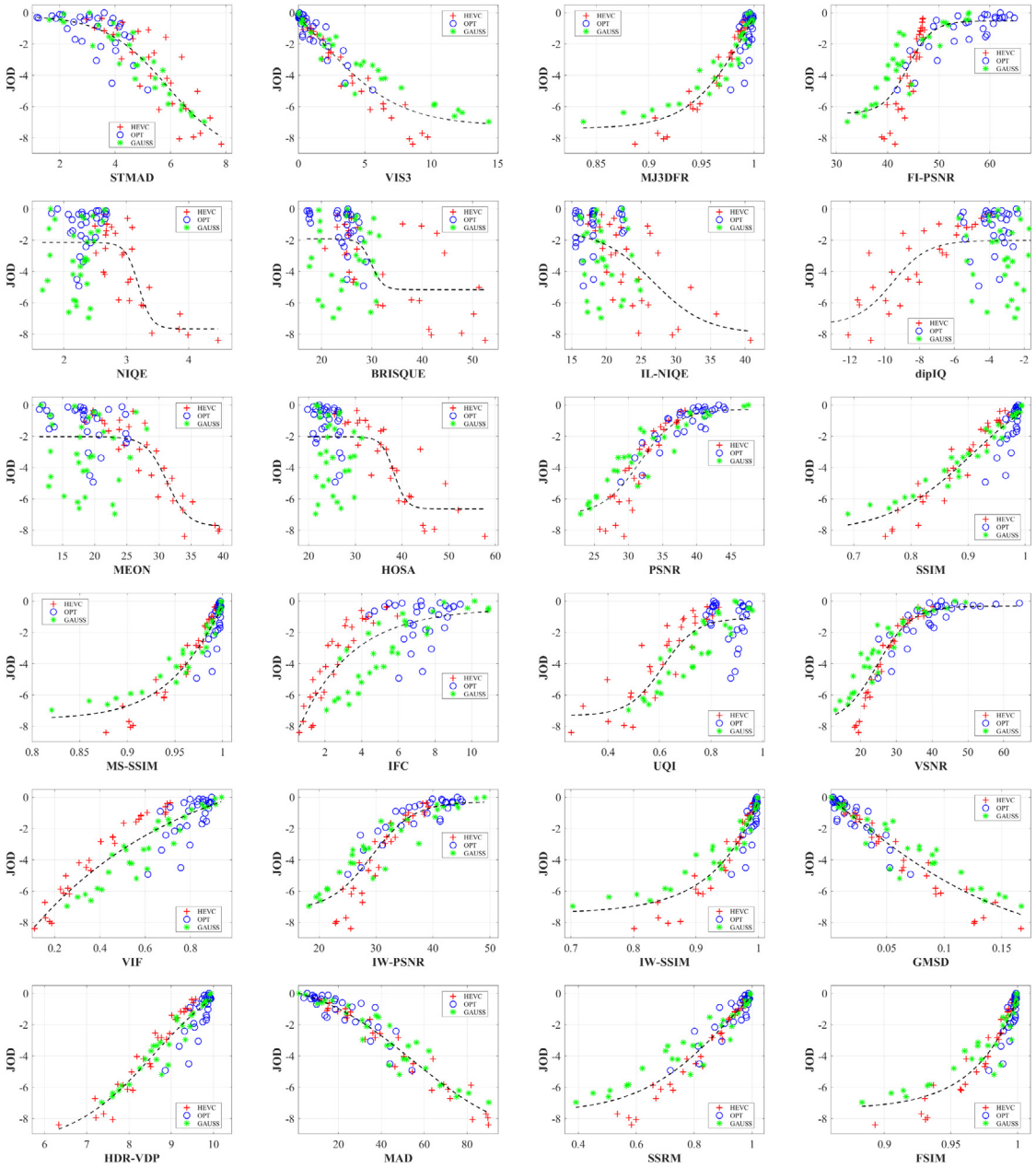


Fig. 11. Scatter plots of Objective metrics vs. JOD (MPI-Lightfield).

4. Analysis of Schulze

The evaluation benchmarks in the previous section showed that none of the examined metrics can significantly outperform all other methods across all data sets. More, the evaluation indices are not in clear agreement about the significance of the results. A pair of metrics with statistically equivalent performance in terms of one evaluation index might be determined to have significantly different performance when deploying another index. Nonetheless, the significance test reveals that there are some metrics with consistently high performance in all data sets. In order to find the most reliable objective metrics for LF quality assessment, we adopted the voting system developed by Markus Schulze for decision-making, which selects the best performing candidates based on the preferential ordering of candidates. The Schulze method [67,68] has been used as a voting analysis mechanism by various organizations and internet decision support systems [68].

The input of the Schulze method is an ordered list of candidates representing the preferences of each voter, where ties are also permitted (*i.e.* candidates with identical rank). Here, the five evaluation indices (*i.e.* LCC, SRCC, RMSE, OR, and AUC) are considered as voters who specified their preferences using the statistical significance test. Comparing a pair of candidates (objective metrics) and a voter (an evaluation index), one candidate is preferred over the other if the performance superiority, determined by that evaluation index, is statistically significant. An insignificant difference (statistical equivalence) implies the same preference over two candidates. In total, 20 voters on four data sets provide their preferences over 24 objective metric candidates and the results are analyzed by the Schulze decision-making mechanism.

The Schulze method is defined based on finding the strength of path between a pair of candidates; Let $N[\alpha, \beta]$ be the number of times that candidate α is preferred to candidate β. Then, a path from candidate α

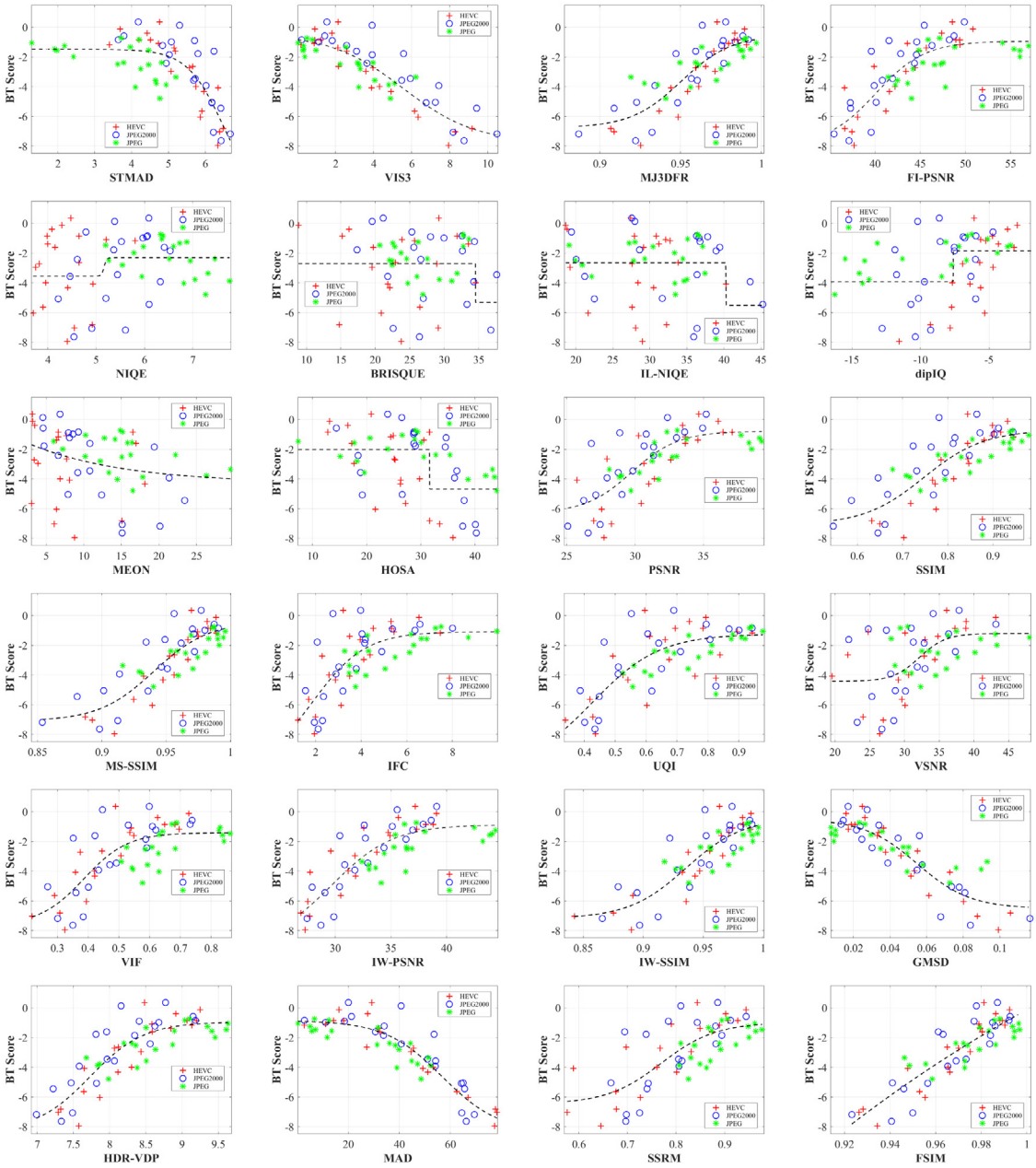


Fig. 12. Scatter plots of Objective metrics vs. BT scores (SMART).

to β is a sequence of candidates $p(1), \ldots, p(n)$ in which:

$$\alpha \equiv p(1), \beta \equiv p(n) \;\; where \;\; n \in \mathbb{N}, \;\; 1 < n < \infty \tag{11}$$

There is at least one direct path ($n = 2$) with strength S_d between a pair of candidates and there can also be indirect paths between two candidates (n>2). The strength of an indirect path $S_{ind}[\alpha, \beta]$ is the strength of the weakest link in the path: $S_{ind}[\alpha, \beta] = min\{N[p(i), p(i+1)]|i = 1, \ldots, (n-1)\}$. Considering direct and indirect paths, the strength of the strongest path from α to β, $S[\alpha, \beta]$ is the maximum strength of the path(s) linking these two candidates: $S[\alpha, \beta] = max\{S_d[\alpha, \beta], S_{ind}[\alpha, \beta]\}$. Finding the strongest path is a well-known problem in graph theory and can be calculated using the Floyd–Warshall algorithm [69].

By applying the Schulze method, we find the strength of path between each pair of candidates. The sum of strength values from each candidate against all other candidates is computed and summarized in Table 7 (in ascending order). As shown in the table, three metrics IW-SSIM, GMSD, and MAD have higher overall strength compared to other methods implying the preference of these three methods. Two widely used methods, PSNR and SSIM, are in rank 14th and 13th of the list, respectively. Fig. 8 presents strength values between a row metric and a column metric for top 14 methods. A candidate α is a winner over β if S[α, β]>S[β, α].

Table 7
Sum of Schulze strength values for each objective quality metric.

NIQE	**MEON**	**BRISQUE**	**dipIQ**	**IL-NIQE**	**HOSA**	**FI-PSNR**	**STMAD**
0	12	18	18	24	33	85	91
VSNR	**UQI**	**PSNR**	**SSIM**	**IFC**	**MJ3DFR**	**IW-PSNR**	**SSRM**
112	135	149	169	172	186	187	194
VIF	**HDR-VDP**	**Vis3**	**MS-SSIM**	**FSIM**	**IW-SSIM**	**GMSD**	**MAD**
198	199	202	209	223	242*	273*	287*

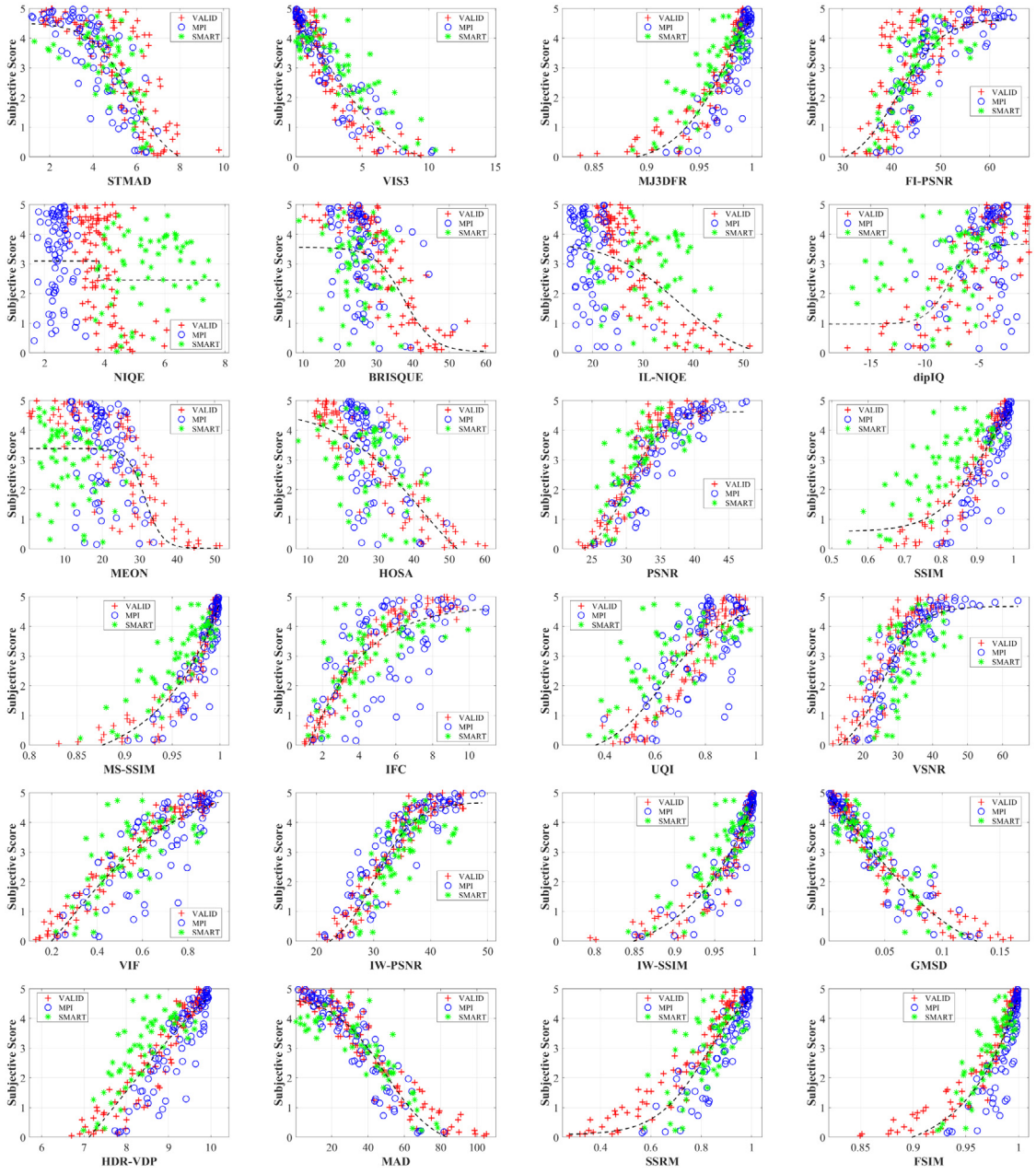


Fig. 13. Scatter plots of Objective metrics vs. Subjective scores (Integrated data set).

5. Metric fusion

In order to further improve the accuracy of the quality assessment, a metric fusion strategy is employed. Since objective metrics might consider diverse visual properties for quality prediction, relying only on a single quality metric might not guarantee the best performance under various distortion types and severity levels. The performance variations of different metrics can be better understood by comparing their results over four data sets in Tables 2 to 6. The idea of metric fusion is to combine the most effective single objective metrics to achieve a more robust quality estimator. Oszut [70] proposed a decision fusion method which computes a weighted product of FR methods for quality assessment. The weights were estimated by solving an optimization problem using a genetic algorithm. Barri et al. [71] focused on improving learning approaches for metric fusion and developed an interpretable machine learning system namely Locally Adaptive Fusion (LAF). Choudhury et al. [72] proposed a fusion strategy for combining the quality metrics for HDR IQA.

The ten best IQA metrics selected by the Schulze method (Table 7) were considered for metric fusion. The aim is to find the best combination of these metrics that can provide the highest performance. Assuming a set of n single quality metric outputs $T = \{q_1, q_2, \dots, q_n\}$, $C(T)$ is a combined metric derived by weighted aggregation of the single metrics. The weights can be computed by solving an optimization problem that minimizes $\|C(T) - MOS\|$. Here we used Support Vector Regression (SVR) to learn the weights. Suppose a given set of m training data $(\mathbf{q_i}, \mathbf{y_i}), i = \{1, 2, \dots, m\}$ where $\mathbf{q_i} = \{q_{i1}, q_{i2}, \dots, q_{in}\}$ and $\mathbf{y_i} = MOS_i$, the SVR [73] aims to find a hyper-plane $h(\mathbf{q}) = \mathbf{w}.\mathbf{q} + \mathbf{b}$ where $\|h(\mathbf{q_i}) - \mathbf{y_i}\| < \epsilon$. The weight parameters $\mathbf{w}$ and bias $\mathbf{b}$ are estimated by applying SVR on the training data. Then, a combined score can be obtained by aggregating single quality metrics with suitable weights. We used 10-fold cross validation to select the best training model.

Using low-performance metrics in a fusion framework can degrade the final combined metrics. Moreover, including multiple high-performance metrics with similar characteristics might not have an

Table 8
Median Performance of single and combined objective quality metrics over 20 iterations on the integrated data set. HDR-VDP (Q_1), Vis3 (Q_2), IW-PSNR (Q_3), VIF (Q_4), MS-SSIM (Q_5), SSRM (Q_6), FSIM (Q_7), IW-SSIM (Q_8), MAD (Q_9), GMSD (Q_{10}).

Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	LCC	SRCC	RMSE	OR
✓	–	–	–	–	–	–	–	–	–	0.893	0.892	0.735	0.725
–	✓	–	–	–	–	–	–	–	–	0.911	0.893	0.667	0.730
–	–	✓	–	–	–	–	–	–	–	0.919	0.905	0.646	0.695
–	–	–	✓	–	–	–	–	–	–	0.898	0.878	0.728	0.695
–	–	–	–	✓	–	–	–	–	–	0.901	0.894	0.709	0.675
–	–	–	–	–	✓	–	–	–	–	0.891	0.885	0.735	0.755
–	–	–	–	–	–	✓	–	–	–	0.904	0.893	0.711	0.695
–	–	–	–	–	–	–	✓	–	–	0.926	0.915	0.610	0.675
–	–	–	–	–	–	–	–	✓	–	0.939	0.915	0.566	0.690
–	–	–	–	–	–	–	–	–	✓	0.942	0.934	0.545	0.635
–	–	–	–	–	–	–	–	✓	✓	0.944	0.936	0.523	0.630
–	–	–	–	–	–	✓	–	✓	✓	0.948	0.937	0.519	0.620
–	–	–	✓	–	–	✓	–	✓	✓	0.953	0.938	0.507	0.610
–	–	–	✓	–	–	✓	✓	✓	✓	0.953	0.938	0.503	0.605

added value and cannot improve performance. Therefore, it is important to choose a subset of the metrics with the most significant effect. To this end, we started with an empty set T' and iteratively added the most significant metrics in terms of LCC improvement. The first metric is the one with the highest LCC value (GMSD). Then in each iteration, a new metric is chosen to be added to the set T' such that $LCC(C(T'), MOS)$ is maximized. The combined metric is trained on the integrated data set in which 60% of images are used for training and the remaining 40% of images are used as test subset. The train-test split is performed 20 times and the median performance across 20 iterations is reported.

Table 8 presents the median performance of the ten selected single metrics as well as the combined methods over 20 iterations. Performance improvement has been achieved by adding new metrics in each step and the highest performance was observed by using five metrics while including other remaining methods could not further improve the results. Fig. 9 shows the distribution of LCC values obtained in 20 iterations where median, minimum, maximum, 25th and 75th percentiles of LCC values are indicated in the form of boxplots. The combined metrics generally indicate better performance compared to the single metrics.

6. Conclusion

This paper presents a comprehensive benchmark of different objective metrics for quality assessment of LF contents. Three LF data sets with subjective scores were considered in which the subjective scores of the three data sets were aligned using the INLSA framework to build an integrated data set including 250 LFs. Different types of metrics including FR and NR IQA approaches, 3D IQA methods, and video metrics were used for the quality prediction task. The accuracy of the objective methods was compared against subjective scores of each data set using different correlation indices. Moreover, the statistical significance of the comparisons was reported. Using the significance test results, we adopted the Schulze decision-making framework to choose the most reliable objective metrics for the LF application. Finally, several quality assessment methods were combined using the SVR framework to achieve higher performance.

Based on our extensive experiments followed by the analysis of Schulze, we concluded that the MAD, IW-SSIM, and GMSD IQA metrics have the highest performance among all benchmark metrics for LF quality assessment. The SSIM and PSNR methods, that are widely used in the literature, are only in the rank 13th and 14th, respectively. The NR metrics are not generally performing well on LF data sets compared to other metric types and better algorithms are needed. In the metric fusion framework, the highest performance (in terms of LCC and SRCC) is achieved by fusing five single quality metrics while adding more methods could not further improve the accuracy of predictions.

CRediT authorship contribution statement

Saeed Mahmoudpour: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation Writing - original draft, Visualization. **Peter Schelkens:** Conceptualization, Methodology, Formal analysis, Resources, Data curation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research work is funded by the research project imec ICON ILLUMINATE, Belgium HBC.2018.0201.

Appendix. Supplementary

A.1. Scatter plots of objective metrics vs. MOS (VALID)

See Fig. 10.

A.2. Scatter plots of objective metrics vs. JOD (MPI-lightfield)

See Fig. 11.

A.3. Scatter plots of objective metrics vs. BT scores (SMART)

See Fig. 12.

A.4. Scatter plots of objective metrics vs. subjective scores (integrated data set)

See Fig. 13.

References

[1] A. Gershun, The light field, J. Math. Phys. 18 (1–4) (1939) 51–151, http://dx.doi.org/10.1002/sapm193918151.
[2] E.H. Adelson, J.R. Bergen, The plenoptic function and the elements of early vision, in: Computational Models of Visual Processing, MIT Press, 1991, pp. 3–20.
[3] M. Levoy, P. Hanrahan, Light field rendering, in: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '96, ACM, 1996, pp. 31–42.
[4] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, Y. Liu, Light field image processing: An overview, IEEE J. Sel. Top. Sign. Proces. 11 (7) (2017) 926–954.
[5] ISO/IEC JTC1/SC29/WG1, JPEG Pleno Call for Proposals on Light Field Coding, wG1N74014, in: 74th JPEG Meeting, Geneva, Switzerland, 2017.
[6] P. Schelkens, Z.Y. Alpaslan, T. Ebrahimi, K.-J. Oh, F. Pereira, A. Pinheiro, I. Tabus, Z. Chen, JPEG Pleno: a standard framework for representing and signaling plenoptic modalities, in: A.G. Tescher (Ed.), Proceedings of SPIE Optics + Photonics: Applications of Digital Image Processing XLI, in: Applications of Digital Image Processing XLI, vol. 10752, SPIE, United States, 2018, p. 107521P, http://dx.doi.org/10.1117/12.2323404.
[7] G.J. Sullivan, J.M. Boyce, Y. Chen, J.-R. Ohm, C.A. Segall, A. Vetro, Standardized extensions of high efficiency video coding (HEVC), IEEE J. Sel. Top. Sign. Proces. 7 (6) (2013) 1001–1016, http://dx.doi.org/10.1109/JSTSP.2013.2283657, Conference Name: IEEE Journal of Selected Topics in Signal Processing.
[8] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, Y.-K. Wang, Overview of the multiview and 3D extensions of high efficiency video coding, IEEE Trans. Circuits Syst. Video Technol. 26 (1) (2016) 35–49, http://dx.doi.org/10.1109/TCSVT.2015.2477935.

[9] S. Rogge, D. Bonatto, J. Sancho, R. Salvador, E. Juarez, A. Munteanu, G. Lafruit, MPEG-I depth estimation reference software, in: 2019 International Conference on 3D Immersion, IC3D, (ISSN: 2379-1780) 2019, pp. 1–6, http://dx.doi.org/10.1109/IC3D48390.2019.8975995.
[10] K. Honauer, O. Johannsen, D. Kondermann, B. Goldluecke, A dataset and evaluation methodology for depth estimation on 4D light fields, in: Asian Conference on Computer Vision, Springer, 2016, pp. 19–34.
[11] M. Rerabek, T. Ebrahimi, New light field image dataset, in: 8th International Conference on Quality of Multimedia Experience, QoMEX, 2016, pp. 1–2.
[12] The (New) stanford light field archive, URL http://lightfield.stanford.edu/.
[13] P. Paudyal, F. Battisti, M. Sjostrom, R. Olsson, M. Carli, Towards the perceptual quality evaluation of compressed light field images, IEEE Trans. Broadcast, 63 (3) (2017) 507–522, http://dx.doi.org/10.1109/tbc.2017.2704430.
[14] V. Kiran Adhikarla, M. Vinkler, D. Sumin, R.K. Mantiuk, K. Myszkowski, H.-P. Seidel, P. Didyk, Towards a quality metric for dense light fields, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[15] L. Shi, W. Zhou, Z. Chen, J. Zhang, No-reference light field image quality assessment based on spatial-angular measurement, IEEE Trans. Circuits Syst. Video Technol. (2019) 1, http://dx.doi.org/10.1109/tcsvt.2019.2955011.
[16] Z. Luo, W. Zhou, L. Shi, Z. Chen, No-reference light field image quality assessment based on micro-lens image, in: 2019 Picture Coding Symposium, PCS, 2019, pp. 1–5.
[17] Y. Tian, H. Zeng, J. Hou, J. Chen, K. Ma, Light field image quality assessment via the light field coherence, IEEE Trans. Image Process. 29 (2020) 7945–7956.
[18] W. Zhou, L. Shi, Z. Chen, J. Zhang, Tensor oriented no-reference light field image quality assessment, IEEE Trans. Image Process. 29 (2020) 4070–4084, http://dx.doi.org/10.1109/tip.2020.2969777.
[19] C. Perra, P. Astola, E.A. da Silva, H. Khanmohammad, C. Pagliari, P. Schelkens, I. Tabus, Performance analysis of JPEG pleno light field coding, in: Applications of Digital Image Processing XLII, SPIE, 2019.
[20] H. Amirpour, A.M. Pinheiro, M. Pereira, M. Ghanbari, Reliability of the most common objective metrics for light field quality assessment, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2019, pp. 2402–2406.
[21] I. Viola, T. Ebrahimi, VALID: Visual quality assessment for light field images dataset, in: Tenth International Conference on Quality of Multimedia Experience, QoMEX, IEEE, 2018, http://dx.doi.org/10.1109/qomex.2018.8463388.
[22] T. Ebrahimi, S. Foessel, F. Pereira, P. Schelkens, JPEG Pleno: Toward an efficient representation of visual reality, IEEE Trans. Multimedia 23 (4) (2016) 14–20, http://dx.doi.org/10.1109/MMUL.2016.64.
[23] P. Schelkens, P. Astola, E.A.B.D. Silva, C. Pagliari, C. Perra, I. Tabus, O. Watanabe, JPEG Pleno light field coding technologies, in: Applications of Digital Image Processing XLII, International Society for Optics and Photonics, 2019, http://dx.doi.org/10.1117/12.2532049.
[24] ISO/IEC JTC 1/SC29/WG1N84065: Information technology - JPEG Pleno Plenoptic image coding system - Part 2: Light Field Coding, 2019.
[25] M.H. Pinson, S. Wolf, An objective method for combining multiple subjective data sets, in: Visual Communications and Image Processing, Vol. 5150, International Society for Optics and Photonics, 2003, pp. 583–592.
[26] J. Liu, T. Malzbender, S. Qin, B. Zhang, C.-A. Wu, J. Davis, Dynamic mapping for multiview autostereoscopic displays, in: Stereoscopic Displays and Applications XXVI, Vol. 9391, International Society for Optics and Photonics, 2015, p. 93911I.
[27] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, Y.-K. Wang, Overview of the multiview and 3D extensions of high efficiency video coding, IEEE Trans. Circuits Syst. Video Technol. 26 (1) (2015) 35–49.
[28] R.A. Bradley, M.E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons, Biometrika 39 (3/4) (1952) 324, http://dx.doi.org/10.2307/2334029.
[29] L.L. Thurstone, A law of comparative judgement, Psychol. Rev. 34 (4) (1927) 273–286.
[30] J. Handley, Comparative analysis of bradley-terry and thurstone-mosteller paired comparison models for image quality assessment, in: PICS, 2001.
[31] M. Perez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, R.K. Mantiuk, From pairwise comparisons and rating to a unified quality scale, IEEE Trans. Image Process. 29 (2020) 1139–1151, http://dx.doi.org/10.1109/tip.2019.2936103.
[32] P. Hanhart, L. Krasula, P.L. Callet, T. Ebrahimi, How to benchmark objective quality metrics from paired comparison data? in: 2016 Eighth International Conference on Quality of Multimedia Experience, QoMEX, IEEE, 2016, http://dx.doi.org/10.1109/qomex.2016.7498960.
[33] E. Zerman, V. Hulusic, G. Valenzise, R.K. Mantiuk, F. Dufaux, The relation between MOS and pairwise comparisons and the importance of cross-content comparisons, Electron. Imaging 2018 (14) (2018) 1–6, http://dx.doi.org/10.2352/issn.2470-1173.2018.14.hvei-517.
[34] Y. Li, M. Sjöström, R. Olsson, U. Jennehag, Scalable coding of plenoptic images by using a sparse set and disparities, IEEE Trans. Image Process. 25 (1) (2015) 80–91.
[35] S. Zhao, Z. Chen, Light field image coding via linear approximation prior, in: IEEE International Conference on Image Processing, ICIP, IEEE, 2017, pp. 4562–4566.
[36] W. Ahmad, R. Olsson, M. Sjöström, Interpreting plenoptic images as multi-view sequences for improved compression, in: IEEE International Conference on Image Processing, ICIP, IEEE, 2017, pp. 4557–4561.
[37] I. Tabus, P. Helin, P. Astola, Lossy compression of lenslet images from plenoptic cameras combining sparse predictive coding and JPEG 2000, in: IEEE International Conference on Image Processing, ICIP, IEEE, 2017, pp. 4567–4571.
[38] I. Viola, T. Ebrahimi, ISO/IEC JTC 1/SC29/WG1M76073: Cross-correlation between subjective tests for JPEG Pleno, 2017.
[39] W. Xue, L. Zhang, X. Mou, A.C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, IEEE Trans. Image Process. 23 (2) (2013) 684–695.
[40] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386.
[41] H.R. Sheikh, A.C. Bovik, Image information and visual quality, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3, IEEE, 2004, pp. III–709.
[42] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, Vol. 2, IEEE, 2003, pp. 1398–1402.
[43] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, J. Electron. Imaging 19 (1) (2010) 011006.
[44] A. Ahar, A. Barri, P. Schelkens, From sparse coding significance to perceptual quality: A new approach for image quality assessment, IEEE Trans. Image Process. 27 (2) (2017) 879–893.
[45] P.V. Vu, D.M. Chandler, ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices, J. Electron. Imaging 23 (1) (2014) 013016.
[46] R. Mantiuk, K.J. Kim, A.G. Rempel, W. Heidrich, HDR-VDP-2, ACM Trans. Graph. 30 (4) (2011) 1–14, http://dx.doi.org/10.1145/2010324.1964935.
[47] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, IEEE Trans. Image Process. 20 (5) (2010) 1185–1198.
[48] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[49] H.R. Sheikh, A.C. Bovik, G. De Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. 14 (12) (2005) 2117–2128.
[50] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Lett. 9 (3) (2002) 81–84.
[51] D.M. Chandler, S.S. Hemami, VSNR: A wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process. 16 (9) (2007) 2284–2298.
[52] A. Mittal, R. Soundararajan, A.C. Bovik, Making a “completely blind” image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2012) 209–212.
[53] M.A. Saad, A.C. Bovik, C. Charrier, Blind image quality assessment: A natural scene statistics approach in the DCT domain, IEEE Trans. Image Process. 21 (8) (2012) 3339–3352.
[54] L. Zhang, L. Zhang, A.C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Trans. Image Process. 24 (8) (2015) 2579–2591.
[55] K. Ma, W. Liu, T. Liu, Z. Wang, D. Tao, dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs, IEEE Trans. Image Process. 26 (8) (2017) 3951–3964.
[56] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, W. Zuo, End-to-end blind image quality assessment using deep neural networks, IEEE Trans. Image Process. 27 (3) (2017) 1202–1213.
[57] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, D. Doermann, Blind image quality assessment based on high order statistics aggregation, IEEE Trans. Image Process. 25 (9) (2016) 4444–4457.
[58] M.-J. Chen, C.-C. Su, D.-K. Kwon, L.K. Cormack, A.C. Bovik, Full-reference quality assessment of stereopairs accounting for rivalry, Signal Process., Image Commun. 28 (9) (2013) 1143–1155.
[59] Y.-H. Lin, J.-L. Wu, Quality assessment of stereoscopic 3D image compression by binocular integration behaviors, IEEE Trans. Image Process. 23 (4) (2014) 1527–1542.
[60] P.V. Vu, C.T. Vu, D.M. Chandler, A spatiotemporal most-apparent-distortion model for video quality assessment, in: 2011 18th IEEE International Conference on Image Processing, IEEE, 2011, pp. 2505–2508.
[61] Recommendation ITU-R BT.500-14 - Methodology for the Subjective Assessment of the Quality of Television Images, International Telecommunication Union, 2019.
[62] Recommendation ITU-R P.1401 - Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models, International Telecommunication Union, 2020.
[63] L. Krasula, K. Fliegel, P. Le Callet, M. Klíma, On the accuracy of objective image and video quality models: New methodology for performance evaluation, in: Eighth International Conference on Quality of Multimedia Experience, QoMEX, IEEE, 2016, pp. 1–6.
[64] E. Zerman, G. Valenzise, F. Dufaux, An extensive performance evaluation of full-reference HDR image quality metrics, Qual. User Exp. 2 (1) (2017) http://dx.doi.org/10.1007/s41233-017-0007-4.

[65] K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The balanced accuracy and its posterior distribution, in: 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 3121–3124.
[66] J. Hanley, B. McNeil, A method of comparing the area under two ROC curves derived from the same cases, Radiology 148 (839–843) (1983) 71.
[67] M. Schulze, A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method, Soc. Choice Welf. 36 (2) (2011) 267–303.
[68] M. Schulze, The Schulze method of voting, 2018, arXiv preprint arXiv:1804.02973.
[69] R.W. Floyd, Algorithm 97: shortest path, Commun. ACM 5 (6) (1962) 345.
[70] M. Oszust, Decision fusion for image quality assessment using an optimization approach, IEEE Signal Process. Lett. 23 (1) (2015) 65–69.
[71] A. Barri, A. Dooms, B. Jansen, P. Schelkens, A locally adaptive system for the fusion of objective quality measures, IEEE Trans. Image Process. 23 (6) (2014) 2446–2458.
[72] A. Choudhury, S. Daly, Combining quality metrics using machine learning for improved and robust HDR image quality assessment, Electron. Imaging 2019 (10) (2019) 307–1–307–7.
[73] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.